به نام خدا

تمرین ۱، گزارش مربوط به دیتاست ۱

علیرضا نجاتی، ۹۸۲۲۲۱۰۴

مقدمه:

این دیتاست مربوط به اطلاعات انواع مختلف تلفن همراه می باشد که در چهار کلاس قیمتی مختلف طبقه بندی شده اند. با توجه به این داده ها می توان به ارتباطی که هر یک از ویژگی های تلفن همراه با کلاس قیمتی خود دارد پی برد. اینکه تغییر هر یک از ویژگی ها چه تاثیری بر کلاس قیمت ودیگر ویژگی ها دارد چیزی است که در این تمرین به آن خواهیم پرداخت. همچنین با بررسی این داده ها مدل هایی تعریف می شوند که به وسیله آن مدل ها کلاس قیمت داده ها را می توان پیش بینی کرد.

توضیحات:

۸ سلول ابتدایی مربوط به اتصال به کگل، دریافت دیتاست، دسترسی کتابخانه های مورد نیاز و ساخت دیتافریم از روی دیتاست می باشد.

۱- **تسک اول:**

تسک اول که پاکسازی داده ها در آن انجام شده است، از دو بخش تشکیل شده است. بخش پیداکردن داده ها پوچ و بخش پیدا کردن داده های پرت. همانطور که در سلول ۱۰ مشخص است، داده ی پوچ در دیتاست ما وجود ندارد.در سلول های ۱۱ و ۱۲ داده های پرت مشخص می شوند که با توجه به ناچیز بودن تعداد آن ها و همچنین زیاد نبودن حجم کل داده ها، از آن ها چشم پوشی کردیم.

۲- تسک دوم:

در تسک دوم دریافت اطلاعات آماری و تحلیلی از دیتاست به همراه مصور سازی صورت گرفته است. سلول ۱۳ اطلاعات آماری خوبی از هر ستون به ما می دهد و در سلول ۱۴ نیز تایپ هر کدام از ویژگی ها مشخص شده است.

در سلول ۱۵ نمودار مربوط به وابستگی ویژگی ها به یکدیگر رسم شده است(Correlation Plot). از این نمودار اینگونه برداشت می شود که کلاس قیمت بیشترین وابستگی را به RAM دارد. با نیروی باتری، طول و عرض نیز این وابستگی وجود دارد. همچنین ویژگی های دوربین جلو و دوربین پشت، G ۳ و ۴ G، طول و عرض موبایل و طول و عرض صفحه نمایش با یکدیگر وابستگی خوبی دارند.

در سلول ۱۶ نموداری برای نشان دادن تعداد داده ها در هر کلاس قیمتی رسم شده است که با توجه به هم اندازه بودن این ستون ها می توان به متوازن بودن داده ها پی برد.
در نمودار های سلول ۱۷ مشخص شده است که برای هر بازه قیمتی مقدار RAM در چه محدوده ای قرار دارد که به این محدوده ها در پایین نمودار اشاره شده است.
نمودار های سلول ۱۸ مربوط به توزیع داده ها برای هر یک از ویژگی های بولین است که برای اکثر این ویژگی های تقریبا به طور مساوی داده ها توزیع شده اند.
در سلول ۱۹ نمودار های مربوط به بازه قیمتی بر حسب هر یک از ۴ ویژگی رم، نیروی باتری، طول و عرض رسم شده اند. مشاهده می شود در ۳ تا از این نمودار ها روند صعودی در کلاس قیمتی ۲ شکسته می شود.
سلول ۲۰ نمودار های جعبه برای ارتباط بازه قیمتی و ویژگی های حافظه داخلی، تعداد هسته، دوربین اصلی و مدت زمان مکالمه را در بر دارد.
در سلول ۲۱ نمودار توزیع کلاس قیمتی برحسب رم و نیروی باتری و در سلول ۲۲ همین نمودار برحسب رم و حافظه داخلی ترسیم شده اند.

۳- تسک سوم:

در تسک سوم ۵ آزمون فرض مطرح شده است که به روش های مختلف بررسی شده اند.
فرض اول این است که دوربین اصلی روی دوربین جلویی تاثیر گذار است که به روش pearson تست شد و جواب آن منفی بود(سلول ۲۳).
فرض دوم تاثیر دوسیم بودن بر روی بازه قیمتی است که به روش ۲_samples_Ttest بررسی شد و جواب آن مثبت بود(سلول ۲۴).
در فرض سوم تاثیرگذاری تعداد هسته ها روی بازه قیمتی به روش chi_squared بررسی شد که جواب آن نیز مثبت شد(سلول ۲۵).
فرض چهارم این است که حافظه داخلی روی نیروی باتری تاثیر دارد. تست این فرض به روش ANOVA انجام شد و نتیجه آن مثبت بود(سلول ۲۶و۲۷).
در نهایت فرض پنجم تاثیر وایفای داشتن یا نداشتن روی بازه قیمتی است که به روش ۲_samples_Ttest تست آن انجام شد و جواب هم مثبت بود(سلول ۲۸).

۴- تسک چهارم:

تسک چهارم مدل سازی برای پیش بینی بازه قیمتی است. قبل از مدلسازی باید ویژگی هدف را از باقی ویژگی ها جدا کرده و سپس جدا سازی داده های train و test صورت گیرد که در سلول های ۲۹ و ۳۰ این جداسازی با نسبت ۸۰ به ۲۰ صورت گرفته است.

در سلول های ۳۱ تا ۳۴ سه گروه از ویژگی ها بر اساس correlation با بازه قیمتی مشخص شده اند که مدل ها را بر روی هر یک از این گروه ها تست کرد ونتایج را بررسی کرد.

مدل اول Naive Bayes یا همان بیز ساده است که یک راهکار ساده برای دسته‌بندی و تعیین روشی برای تشخیص برچسب اشیاء یا نقاط می باشد. دقت این مدل برای مجموعه تمام ویژگی ها برابر با ۰,۸۲۲۵ (سلول ۳۶) و برای مجموعه دوم که شامل ۱۰ ویژگی است برابر با ۰,۸۰۵ (سلول ۳۷) و برای مجموعه سوم که شامل ۵ ویژگی است برابر با ۰,۷۹۷۵ (سلول ۳۸) شد.

در مدل دوم Logestic Regression با تغییر پارامتر ها این کلاس سعی می کنیم به دقت مطلوبی برسیم و بعد مدل را بر روی گروه ها اعمال می کنیم. همانطور که در سلول ۳۹ مشخص است با max_iter پیش فرض که برابر با ۱۰۰ است به دقت ۰,۶۵۲۵ می رسیم. با افزایش این پارامتر به ۱۰۰۰ در سلول ۴۰ دقت به ۰,۷۴۵ می رسد و باز هم با افزایش تعداد تکرار به ۲۰۰۰ در سلول ۴۱ به دقت نسبتا مطلوب ۰,۷۶۷۵ خواهیم رسید. دقت این مدل برای مجموعه دوم ۰,۶۳ (سلول ۴۳) و برای مجموعه سوم برابر با ۰,۹۵ (سلول ۴۴) بدست می آید.

مدل سوم Random Forest با random_state = 0 که در سلول ۴۵ بررسی شده است، دقتی برابر با ۰,۸۸۵ را به ما می دهد که افزایش این مقدار تا عدد ۲۰ دقت را نیز افزایش می دهد و برابر با ۰,۹۰۲۵ می شود(سلول ۴۶). همچنین افزایش پارامتر n_estimator تا عدد ۱۵۰ نیز موجب افزایش دقت تا ۰,۹۱ می شود(سلول ۴۷). با همین مقادیر و دقت به سراغ مجموعه ویژگی هایمان می رویم که دقت برای مجموعه اول همان ۰,۹۱ (سلول ۴۸) ، برای مجموعه دوم برابر ۰,۹۱۷۵ (سلول ۴۹) و برای مجموعه سوم ۰,۹۲۷۵ (سلول ۵۰) بدست می آید. همچنین در این قسمت اشاره می شود که هر سه مدل از استراتژی OVA استفاده می کنند.

۵- تسک پنجم:

تسک ۵ رسم Confusion Matrix (ماتریس درهم ریختگی) برای هریک از مدل های بالاست. با بررسی این ماتریس ها ملاحظه می شود مدل Random Forest در سلول های ۵۱ تا ۵۳ برای بازه های قیمتی با برچسب ۰،۱،۳ پیش بینی خوب و نسبتا دقیقی دارد اما برای بازه ۲ عملکرد ضعیف تری دارد. مدل logestic Regression در سلول های ۵۴ تا ۵۶ برای کلاس های ۰ و ۳ پیش بینی های دقیقی دارد اما برای پیش بینی کلاس ها ۱ و ۲ ضعیف عمل می کند. مدل Naive Bayes در سلول

های ۵۷ تا ۵۹ عملکرد مشابهی دارد که دلیل این ضعف مدل ها برای کلاس های ۱و۲ در ادامه همین سلول به همراه نمودار (سلول های ۶۰ و ۶۱) ذکر شده است.

۶- تسک ششم:

تسک ۶ که تشخیص متوازن یا نامتوازن بودن داده هاست در سلول ۶۲ بررسی شده است. که باتوجه به نمودار رسم شده مشخص است که داده ها متوازن اند. همچنین به راهکار های هندل کردن داده های نامتوازن نیز اشاره شده است.

۷- تسک هفتم:

در تسک ۷ عملیات scaling با دو روش MinMax و Standard انجام شده است. MinMaxScaler مقادیر همه ستون ها را به بازه ۰ تا ۱ می برد. در سلول ۶۳ این روش انجام ونتیجه آن به نمایش در آمده است. در سلول ۶۴ داده های اسکیل شده را به همان مدل Logestic Regression با max_iter = ۲۰۰۰ دادیم و نتیجه بسیار بهتر از حالت قبل بود که بصورت جدول و ماتریس درهم ریختگی به نمایش درآمد. دقت در این حالت تقریبا ۰,۱۷ افزایش داشت. در سلول ۶۵ روش Standard انجام شده است. نتیجه حاصل از این روش را نیز به همان مدل ذکر شده دادیم و این بار دقت ۰,۲ افزایش داشت به عدد ۰,۹۶۷۵ رسید. همچنین در این سلول ماتریس درهم ریختگی نیز رسم شده است.

۸- تسک نهم:

تسک ۹ انجام روش PCA یا همان آنالیز مولفه اصلی با POV های مختلف است. این روش برای کاهش بعد بکار می رود یعنی آن ویژگی هایی را که ارزش بیشتری فراهم می کنند برای ما استخراج می کند. در سلول ۶۸ PCA با POV = 0.75 انجام شده است که ملاحظه می شود تعداد ویژگی ها به عدد ۲ کاهش پیدا می کند. سپس در سلول ۶۹ داده های train و test جدید حاصل از PCA را به مدل استفاده شده در تسک ۷ می دهیم و نتیجه را به صورت ماتریس درهم ریختگی رسم می کنیم که دقت آن برابر با ۰,۸۱ می شود که حدود ۰,۰۵ افزایش یافته است. سلول های ۷۰ و ۷۱ مشابه بالا است با این تفاوت که POV=0.8 در این حالت نیز تعداد ویژگی ها همان ۲ می شود و دقت نیز همان ۰,۸۱ می ماند. در سلول های ۷۲ و ۷۳ این کار با POV=0.9 انجام می شود. تعداد ویژگی ها ۳ و دقت برابر ۰,۹۶ می شود. در ادامه POV=0.95 در سلول های ۷۴ و۷۵ بررسی می شود که نتیجه شامل ۴ ویژگی و دقت برابر با ۰,۹۵ بدست می آید. در نهایت در سلول های ۷۶ و ۷۷، POV را برابر ۰,۹۹ قرار می دهیم و نتیجه مشابه حالت قبل است. همانطور که مشخص است نتیجه ها بهبود یافته اند اما نسبت به حالت scale شده تفاوت چشمگیری وجود ندارد.

۹- تسک دهم:

در تسک ۱۰ داده ها را نامتوازن می کنیم و سپس دوباره مدل ها رو روی این داده های نامتوازن تست می کنیم. نتیجه حاصل از این نامتوازن سازی در سلول ۷۸ وجود دارد. سلول ۷۹ جداسازی ستون هدف و جداسازی داده های train و test را در بر دارد. در سلول ۸۰ داده های جدید به مدل Random Forest که دقت بهتری داشت داده می شود و نتیجه بصورت ماتریس درهم ریختگی نمایش داده شده است. دقت مدل برابر با ۰,۹۷۷۵ می شود که افزایش تقریبا ۰,۷ را به همراه داشت. برای مدل Logestic Regression با دقت بهبود یافته این مراحل در سلول ۸۱ انجام شده است و دقت ۰,۹۴۵ بدست آمده است یعنی افزایش به مقدار تقریبی ۰,۱۸. در سلول ۸۲ مدل Naive Bayes بررسی شده است که دقت برای این مدل افزایش تقریبی ۰,۱۴ به همراه داشت و به مقدار ۰,۹۶۵ رسید. این نتایج نشان می دهد که نامتوازن کردن داده های این دیتاست موجب بهبود عملکرد مدل ها شده است. در ادامه داده های نامتوازن را به روش Upper Sample تبدیل به داده های متوازن کردیم (سلول ۸۴) و دوباره مدل ها را روی داده های متوازن شده ی جدید تست کردیم. مدل Random Forest در سلول ۸۸ بررسی شد و دقت برابر با ۰,۹۸۱۶ بدست آمد. مدل Logestic Regression در سلول ۸۹ بررسی شد و دقت برابر با ۰,۹۵۶۶ بدست آمد. مدل Naïve Bayes در سلول ۹۰ بررسی شد و دقت برابر با ۰,۹۴۱۶ بدست آمد. همانطور که مشخص است عملکرد دو مدل اول با متوازن کردن داده های جدید کمی بهتر شده است.